Variational Inference

From EM-days, we know, for "any" $q(z)$

$$\log P(x|\theta) \geq \underbrace{E_{q(z)} \log P(x, z|\theta) - E_{q(z)} \log q(z)}$$

can be derived with Jensen's inequality

What is the difference b/w L.H.S & R.H.S ?

$$\log P(x|\theta) - E_{q(z)} \log P(x, z|\theta) + E_{q(z)} \log q(z)$$

$\Downarrow$ same

$$E_{q(z)} \log P(x|\theta) - E_{q(z)} \log P(x, z|\theta) + E_{q(z)} \log q(z)$$

$$E_{q(z)} \log \frac{P(x|\theta)}{P(x, z|\theta)} + E_{q} \log q(z)$$

$$= -\left[E_{q(z)} \log \frac{P(x, z|\theta)}{P(x|\theta)} - E_{q} \log q(z) \right]$$

$$= - E_{q(z)} \log \frac{P(z|x, \theta)}{q(z)}$$

$= KL(q \,||\, p(z|x,\theta))$: Kulback-Leibler divergence b/w q and $P(z|x,\theta)$

Hence

$$\log P(x|\theta) = \underbrace{E_{q(z)} \log \frac{P(x,z|\theta)}{q(z)}}_{\substack{\text{lower-bound} \\ L(q;\theta)}} - \underbrace{E_{q(z)} \log \frac{P(z|x,\theta)}{q(z)}}_{KL(q \,||\, p(z|x,\theta))}$$

$L(q;\theta)$ ← function of q and θ

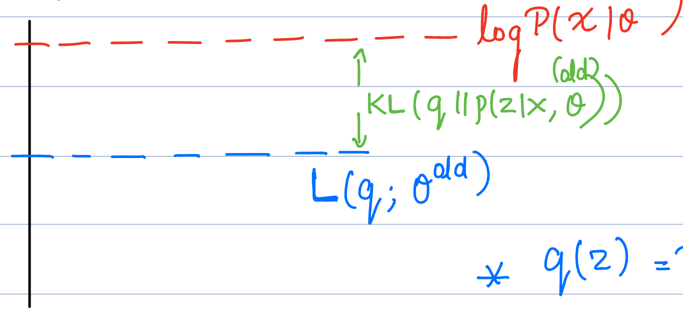
Alternative view of EM :-

* Coordinate ascent between $q(z)$ and θ

In E-step, $\theta = \theta^{(old)}$ and we optimize $L(q;\theta^{(old)})$ w.r.t $q(z)$, $q(z) \longrightarrow p(z|x,\theta^{old})$

In M-step, $q(z) = p(z|x,\theta^{(old)})$ and we optimize $L(p(z|x,\theta^{old});\theta)$ w.r.t to θ

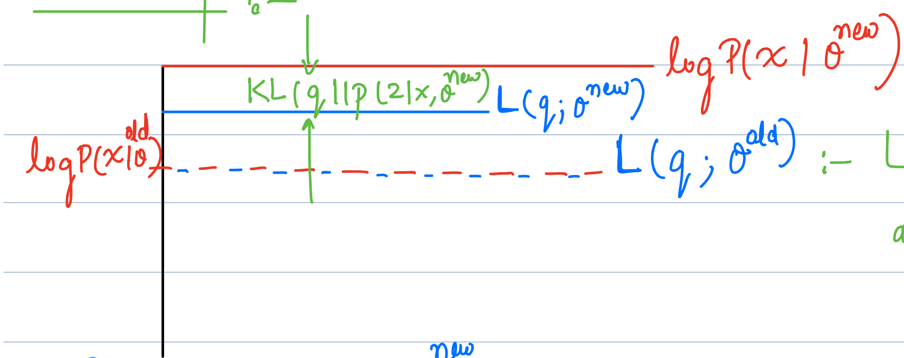
E-Step :-



* $q(z) = p(z|x,\theta^{old})$ is the solution of $\underbrace{\arg\max_{q(z)} L(q;\theta^{old})}$

* In E-step, $\log P(x \mid \theta^{old})$ does not improve as independent of $q(z)$ Imp
* Only $KL(q \,||\, p(z \mid x, \theta)) \longrightarrow 0$, when $q(z) \to p(z \mid x, \theta)$

M-step :-



:- $L(q; \theta^{old}) = \log P(x \mid \theta^{old})$ at start of M-step

* In M-step, $\theta^{new} \longleftarrow \underset{\theta}{\operatorname{argmax}} \; L(q(z) = p(z \mid x, \theta^{old}); \theta)$

* Also moves $\log P(x \mid \theta)$ up, and makes $KL(q \,||\, p(z \mid x, \theta^{new}))$ ≥ 0

* In a Bayesian model, $\theta \sim \text{Prior}(\theta \mid \text{hyper-param})$ ← like any latent variable
* We want $\Pr(\theta \mid x)$
* Essentially, we just want to run the E step above i.e

$$\boxed{\max_{q} L(q; \text{hyper-parameters})} \quad (A)$$

where q is some distribution on latent variables

* $\max_{q} L(q; \text{hyperparameters})$ is optimization in the space of distributions.

* Variational inference deals with this optimization.

* If no constraint is put on the space of distributions, optimal $q = P(\text{latent\_variables} \mid x)$

* $Pr(\text{latent\_variables} \mid x)$ may not be possible to compute analytically. {for various reasons, we have discussed so far}

* Goal :- find a good easy approximation to $Pr(\text{latent\_variables} \mid x)$ by constraining space of "q"

* Two approaches :- a) assume a parametric form of 'q' and optimize $L(q; hyp)$ w.r.t parameters

  b) Assume a factorized 'q' space. (mean-field approximation)

Any of the two approaches will (possibly) leave a non-zero $KL(q \| p(\text{latent variables} \mid x))$

MEAN-FIELD APPROXIMATION

Let $Z = [z_1, \dots, z_M]$ all latent variables

$q(Z) = \prod_{i=1}^{M} q(z_i)$ is the constrained state of distributions. Optimizing $L(q; hyp)$ w.r.t such $q(Z)$ gives the following

$$\boxed{\log q(z_i) = \underset{\prod_{j:j\neq i} q(z_j)}{E} \log P(x, Z) + \text{const}}$$

expectation of $\log P(x, z)$ w.r.t $\prod_{j:j\neq i} q(z_j)$

if $\log = \log_e$ (natural log)

$$q(z_i) = \frac{\exp \underset{j:j\neq i}{E} \ln P(x, z)}{\int \exp \underset{j:j\neq i}{E} \ln P(x, z)\, dz_i} \quad \text{for } i = 1, \dots, N$$

Proof:-

$$L(q\,;\,hyp) = \int \ln P(x, z \mid hyp) \prod_{i=1}^{M} q(z_i)\, dz_1 \ldots dz_M$$

$$- \int \left[\ln \prod_{i=1}^{M} q(z_i)\right] \prod_{i=1}^{M} q(z_i)\, dz_1 \ldots dz_M$$

$$= \int q(z_i) \Big[\overbrace{\int_{j \neq i} \ln P(x, z \mid hyp) \prod_{j: j\neq i} q(z_j)}^{E_{i \neq j} \ln P(x, z \mid hyp)} \Big] dz_i$$

$$- \Big[\int (\ln q(z_i))\, q(z_i)\, dz_i + \underbrace{\sum_{j \neq i} \int [\ln q(z_j)]\, q(z_j)\, dz_j}_{\text{Constant}} \Big]$$

for optimizing w.r.t $q(z_i)$

$$= \int q(z_i)\, E_{j \neq i} \ln P(x, z)\, dz_i$$

$$- \int q(z_i) \ln q(z_i)\, dz_i$$

negative KL b/w $q(z_i)$ and $E_{i \neq j} \ln P(x, z \mid hyp)$

So $\quad q(z_i) = E_{j \neq i} \ln P(x, z) + \text{const}$